文章编号:1001-5078(2024)02-0302-05

·图像与信号处理·

深度学习与图像融合的行人检测算法研究

姜柏军[1],钟明霞[1],林昊昀[2]
(1. 浙江商业职业技术学院,浙江 杭州 310053;2. 首都师范大学数学科学学院,北京 100089)

摘　要:为解决智能辅助驾驶技术中可见光摄像机受光照和气候影响而导致行人目标识别困难的问题。通过研究图像融合技术,结合深度卷积神经网络,实现并改进了一种道路行人目标检测算法。方法是利用多源传感器图像融合技术,采用可见光相机与红外热成像相机融合的策略,以 Faster RCNN 算法为基础,从改进网络结构、特征融合、优化模型训练等方面展开研究,对复杂环境下的行人检测与定位跟踪展开研究,提出一种基于图像融合技术和改进的深度卷积神经网络的道路行人目标检测算法。实验结果表明,该算法对复杂气候环境下行人目标检测提高了检测效率和准确率,增加了智能辅助驾驶汽车的安全性。

关键词:红外热成像;可见光图像;Faster RCNN;深度卷积神经网络;行人目标检测
中图分类号:TN219;TP391.41　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2024.02.021

Research on pedestrian detection algorithm combining deep learning and imaging fusion

JIANG Bo-jun[1], ZHONG Ming-xia[1], LIN Hao-yun[2]
(1. Zhejiang Business College, Hangzhou 310053, China;
2. School of Mathematical Sciences, Capital Normal University, Beijing 100089, China)

Abstract: The aim of this paper is to address the difficulty in pedestrian target recognition in intelligent assisted driving systems due to the influence of light and climate on visible light cameras. A pedestrian target detection algorithm is implemented and improved by studying image fusion techniques in combination with deep convolutional neural networks. Firstly, using multi-source sensor image fusion technology, the strategy of fusing visible light cameras and infrared thermal imaging cameras, based on the Faster RCNN algorithm, a pedestrian target detection algorithm based on infrared thermal imaging technology and improved depth convolutional neural network is proposed. Then, the research is carried out in terms of improving network structure, feature fusion, optimising model training, and so on, and the research is carried out on pedestrian detection and localisation tracking in complex environments. Finally, the experimental results show that this algorithm improves detection efficiency and accuracy for human target detection in complex climate environments, and increases the safety of intelligent assisted driving vehicles.

Keywords: infrared thermal imaging; visible light images; Faster RCNN; deep convolutional neural network; pedestrian target detection

1　引　言

随着中国经济迅速发展,人口众多和城市交通规划的不合理性逐步显现,我国的交通状况日益严重。这导致道路资源日益紧张,交通事故频发。根据资料

基金项目:浙江省教育厅一般科研项目(No. Y202147947)资助。
作者简介:姜柏军(1978 -),男,硕士,讲师,主要从事图像处理和深度学习技术研究。E-mail:bungeejiang@163.com
通讯作者:钟明霞(1982 -),硕士,讲师,主要研究方向为计算机技术、图像处理技术。E-mail:517997106@qq.com
收稿日期:2023-09-19

数据[1]，我国交通事故的致死率高达 27.3 %，居全球之首。同时的调查结果[2]显示，在致死事故中，美国和欧洲国家的死亡人数主要集中在乘车人员，而在中国，超过 60 % 的死亡人数是行人、自行车等交通弱势群体。因为在中国的道路权益分配中，行人和自行车的权益受到机动车严重挤压，人车混合出行的模式导致行人的安全面临严峻挑战。除了各汽车制造商需要逐步建立行人保护安全开发体系外，利用车辆的智能辅助驾驶功能可以有效降低交通事故中的死亡人数[3]。当前，国内外学者主要关注基于两类图像进行行人检测与跟踪的研究：可见光图像和红外图像。然而，可见光摄像头难以应对恶劣天气条件下（如黑夜、弱光、烟雾、雾和蒸汽等）的交通环境。为弥补可见光摄像头的不足，本文提出在汽车传感器套件中加入红外热像仪，把可见光图像和红外图像进行融合，图像融合技术目前也广泛应用于行人目标识别中。将图像融合技术应用于自动驾驶中，可以提升行人的安全性，填补视觉盲区，提供更多决策信息，预防事故和碰撞，同时改善驾驶体验。通过将红外图像的热能分布与可见光图像的视觉特征相结合，可以在夜间和低光条件下更精准地检测和跟踪道路上的障碍物，如行人、车辆等。红外图像与可见光图像的融合提供了更全面的感知能力，从而提升了自动驾驶系统的安全性和鲁棒性。

近年来，为了提升行人检测效果，伴随着新算法的涌现以及硬件的升级，利用深度学习从图像中提取特征并进行行人目标判断的技术逐渐增多，其中包括 R-CNN[4]、YOLO[5]、SSD[6] 等几类主流框架。研究文献表明，可见光图像下的行人检测方法已经相对成熟，但目前涉及可见光和红外图像的行人检测方法尚处于初级阶段，需要克服诸多难题。这些难题主要集中在以下两个方面：（1）受白天和夜间光照变化的影响，可见光和红外图像融合特征在不同光照条件下表现出差异性。（2）目前，基于深度卷积神经网络的行人检测模型常常表现出较高的检测率，但其效率相对较低，未能同时保障实时性和准确性，难以满足辅助驾驶实时检测的需求。当前行人目标检测算法在特定情况下面临着挑战，例如夜间、低能见度和复杂背景等，这些环境条件的影响可能导致行人目标检测的准确性下降，因此需要更为强大的方法来应对上述问题。

2　双模态特征提取与融合

为了克服热成像的局限性，并提高道路目标识别的准确性和可靠性，可以采用图像融合技术。图像融合通过将热图像与可见光图像进行融合，结合它们的优势，从而产生一个融合图像，使得图像中既包含了热能信息，又保留了可见光的颜色和纹理信息。因为单一可见光或红外图像分类器在全天候驾驶环境中无法有效识别在白天和夜间光照环境下存在差异性的行人特征而导致出现漏检情况，本文在基于区域生成网络的可见光与红外图像行人目标检测器的基础上，进行可见光与红外图像双模态特征融合，以优化深度卷积神经网络分类性能，提高行人检测准确率。首先采用双模态区域生成网络即双路深度卷积神经网络分别对可见光图像和红外图像进行特征提取，提取得到的可见光特征与红外特征通过级联融合后输入区域生成网络进行特征分类和回归。双路深度卷积神经网络，均由 5 个卷积层（Conv）和 4 个池化层（Pool）交替堆栈组成。如图 1 所示，可见光图像特征提取模块的卷积层从 Convl-V 到 Conv5-V，池化层从 Pooll-V 到 Pool4-V；红外图像特征提取模块的卷积层从 Convl-I 到 Conv5-I，池化层从 Pooll-I 到 Pool4-I；双模态区域生成网络特征融合模块采用级联融合层（Concat）将可见光特征与红外特征级联在一起，进过融合卷积层（Conv-F）进行融合特征学习后，输出可见光与红外融合特征。

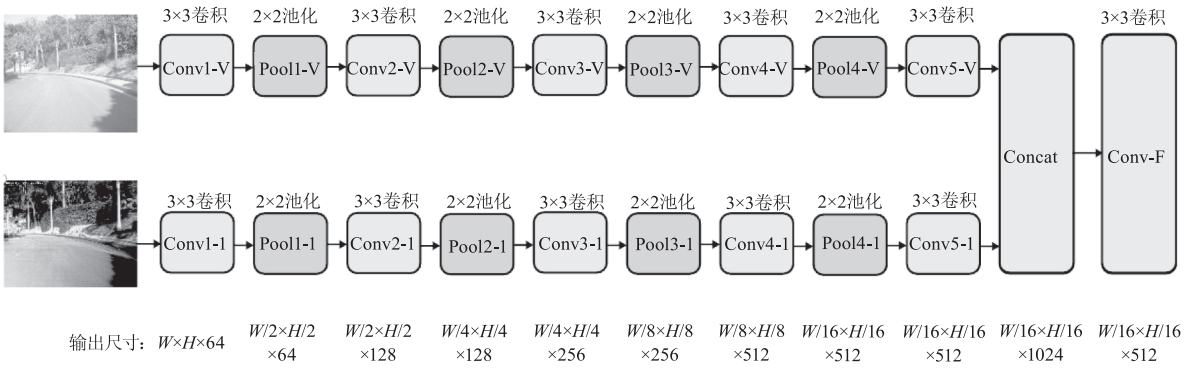

图 1　双模态区域生成网络特征提取与融合模块网络结构图

Fig. 1 Network structure diagram of feature extraction and fusion module for bimodal region generation network

双模态区域生成网络特征提取网络结构采用 VGGl6 深度卷积神经网络架构,所有的卷积层采用 3 ×3 卷积核,所有的池化层采用 2 ×2 池化核,具体的参数设计如表 1 所示。采用 W 和 H 分别表示可见光图像和长波红外图像的长宽像素值。此处,可见光和红外图像输入尺寸均 $W \times H \times 3$,特征融合模块输出的可见光与红外融合特征图尺寸为 $W/16 \times H/16 \times 512$。

3 改进的 Faster RCNN 算法

本文在 Faster RCNN 基础上,针对红外热成像技术[7]通过四种措施来提升 Faster RCNN 在汽车驾驶场景下的行人目标检测性能:①设计特征融合网络,并构建了一种感兴趣候选区域空间金字塔池化网络,以提高算法在汽车驾驶场景的行人目标检测性能;②通过聚类算法对训练数据集中真值框的宽高信息进行聚类,利用聚类结果优化锚设置,挖掘汽车驾驶场景下的先验知识来提升检测算法的检测效率;③采用在线案例挖掘技术优化模型训练;④对网络卷积层中的函数进行改进,并调整训练参数来提高系统分类性能。

3.1 改进网络结构

Faster RCNN 算法中需要先设计特征提取网络,用于特征提取。针对基本算法中存在的问题主要是:①候选框选择机器耗时长;②候选框覆盖面广,重叠区域特征重复计算;③步骤多,过程复杂。原始 RCNN 重复使用深层卷积网络在 2k 个窗口上提取特征,特征提取非常耗时。我们在这里改进了 RCNN 的不足,采用空间金字塔池化网络(图 2)中 SPPNet 将比较耗时的卷积计算对整幅图像只进行一次,之后使用 SPP 将窗口特征图池化为一个固定长度的特征表示。

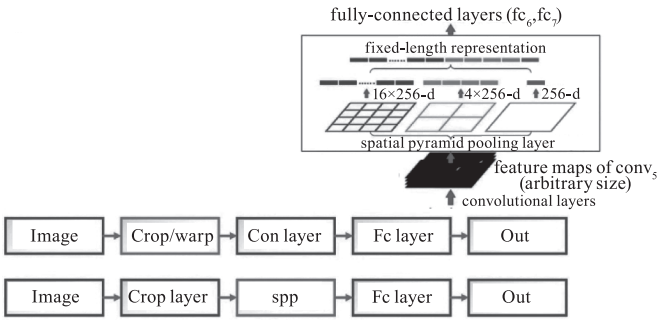

图 2 空间金字塔池化网络

Fig. 2 Spatial pyramid pooling network

对于上图中的网络,Image 是输入图像就是候选区域,Crop/warp 就是对候选区域进行提取,然后将图像 resize 到固定的大小。由于网络中加入这两个操作,使得候选区域会出现扭曲的情况。因此把固定大小的图像输入到卷积神经网中,尽可能特征提取,最后在 FC 层得到输出的特征向量。我们采用同一个卷积神经网,需要保证输入图像大小必须统一。这里把候选区域的提取安排在图像输入的下一个环节,根据不同的候选区域会都采用相同卷积来完成特征提取的过程,导致重复计算,也是 RCNN 网络存在的问题。重新优化在上图底部,加入 spp 层,对于不同尺寸提取不同维度的特征,它会将每一个卷积层的输出固定的通过 SPP 层得到一个 21 维特征,这个 21 维是对应每个 feature map 的,即对每一个通道数(channel),具体维数值 $21 \times c$,就是通过 SPP 层产生固定的输出,再通过 FC 层计算。

3.2 模型训练

Faster RCNN 是两个阶段的检测器,模型训练要分两个部分:一是训练 RPN 网络,二是训练后面的分类网络。为了能够说明模型训练过程,需要明确提及的两个网络。分别是:RPN 网络(共享 conv 层 +RPN 特有层);Faster RCNN 网络(共享 conv 层 + Faster RCNN 特有层)。训练的步骤过程如下:

①先用 ImageNet 的预训练权重初始化 RPN 网络的共享 conv 层,再训练 RPN 网络。训练完成,即更新了共享 conv 层和 RPN 特有层的权重;

②根据训练好的 RPN 网络获取 proposals;

③再使用 ImageNet 的预训练权重初始化 Faster RCNN 网络的贡献 conv 层,然后训练 Faster RCNN 网络。随着训练完成,再次更新共享 conv 层和 Faster RCNN 特有层的权重;

④利用步骤③训练好的共享 conv 层和步骤①训练好的 RPN 特有层来初始化 RPN 网络,进行第二次训练 RPN 网络。这里固定共享 conv 层的权重,保持权重不变,只训练 RPN 特有层的权重;

⑤根据训练好的 RPN 网络获取 proposals;

⑥继续使用步骤③训练好的共享 conv 层和步骤③训练好的 Faster RCNN 特有层来初始化 Faster RCNN 网络,再次训练 Faster RCNN 网络。在这里,固定 conv 层,只保留 fine tune 特有部分。模型训练过程如图 3 所示。

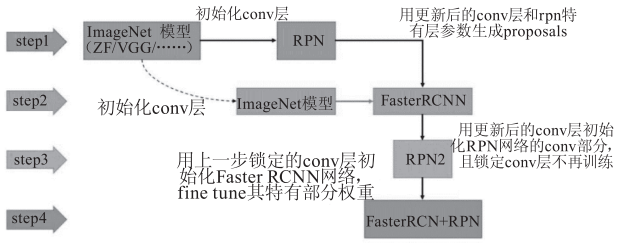

图 3　模型训练步骤过程

Fig. 3 Process of model training steps

3.3　改进函数

在卷积层候选框训练提取网络的时,把锚分为两类,选中目标的锚标记是正样本(positive),未选中目标的锚标记是负样本(negative)。正样本就是通过锚和真值相交的情况来定义,通过两种方式实现。对于每个真值,存在两种情况,所有锚与这个真值要么相交,要么不相交。相交方式中:和这个真值有最大交并比的那个错误标记为正样本;与这个真值的交并比在大于 0.7 时,那些锚也标记为正样本。负样本就是与所有真值的交并比在小于 0.3 时的锚。除了以上,其余的锚无需标记,在整个模型训练过程中不使用。

根据正负样本的定义,给出 RPN 损失函数的公式(1)所示:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \tag{1}$$

对于以上公式,实际由两部分组成。第一部分计算分类误差,第二部分计算回归误差。计算分类误差时,p_i 是一个 anchor box 值为正值的概率,$p^*{}_i$ 是 anchor box 的真实数据,取值为 0 或 1,这里研究采用二分类 log loss,$\sum$对所有 anchor box 的 log loss 累加求和;计算回归误差时,t_i 表示预测的 anchor box 位置,$t^*{}_i$ 表示与 anchor box 相关的真实数据的位置,L_{reg} 实际上是计算 t_i 和 $t^*{}_i$ 的位置差,也被称为平滑 L_1,在对所有的 anchor box 计算结果的误差累加求和时,仅仅计算正值类 anchor box 的累加和。关于系数部分,N_{cls} 的取值为最小批次中 anchor box 的数量,即 $N_{cls}=256$;N_{reg} 为一张图对应的 anchor 的数量,数值约等于 2400;为了确保函数两部分 loss 前的系数最大程度相同,设置 $\lambda=10$。

4　实验结果与分析

为了验证实验效果,本文测试数据库采用 2019 年 8 月 FLIR 公司推出的免费用于算法训练的 FLIR Thermal Starter 数据集 V1.3。数据是由安装在车上的 RGB 相机和热成像相机获取的。数据集总共包含 14452 张红外图像,其中 10228 张来自多个短视频;4224 张来自一个长为 144 s 的视频;数据集图像包括 5 种目标分类:行人、狗、机动车、自行车及其他车辆。该数据集使用 MSCOCO labelvector 进行标注,提供了带注释的热成像数据集和对应的无注释 RGB 图像(图 4),数据集文件格式包括五种:(1)14 位 TIFF 热图像(无 AGC);(2)8 位 JPEG 热图像(应用 AGC),图像中未嵌入边界框;(3)8 位 JPEG 热图像(应用 AGC),图像中嵌入边界框便于查看;(4)RGB-8 位 JPEG 图像;(5)注释:JSON(MSCOCO 格式)。

本文在改进的空间金字塔网络结构中,设计了 6 个 anchor 来覆盖整个输入的图片,anchor 的长宽比例设置为[1∶1,1∶2]。通过大量的实验数据得出,采用这个参数设置算法效果相对最好。实验中,我们先对红外图像做了预处理,即红外图像和可见光图像做的融合处理,如图 4 所示。本文采用的算法实现道路行人目标识别的效果图,如图 5 所示。

(a)红外　(b)可见光　(c)融合

图 4　道路三种图像效果图

Fig. 4 Three image renderings of roads

图 5　行人识别效果图

Fig. 5 Pedestrian recognition effect

通过算法的设计在 python 程序中的实现,经过模型训练。我们做出如下分析:①比较 2 分类和 3 分类道路识别:3 分类是背景,行人,骑自行车和骑摩托车的人,通过误差分析,行人和骑车的人经常混

淆,然后说明了可以把行人和骑车的人合并在一起的理由,进行了 2 分类测试,效果显然比三分类好。②卷积通道调整:在测试识别过程中发现了一些顽固的反例,这些样本是由灯光反射或车辆灯光系统造成的。在训练和测试中为了减轻这些反例的影响,应用均值减去法对样本数据进行预处理。此外,为防止梯度爆炸,在训练过程中当测试损失率连续 3 代不再提高的时候将学习率减半。对比了卷积层滤波器个数为 30 ~ 18,25 ~ 15,20 ~ 12 时的 2 分类结果。得到个数为 25 ~ 15 时 2 分类结果最佳,测试准确率 93. 22 %,训练损失率为 1. 07 %。③使用自学习 softmax 分类器微调:准确率由 93. 22 % 上升到 94. 49 %,平均识别时间为 0. 07 ms。

本文从 FLIR Thermal Starter 数据集中选择用于测试的实验红外热图像 600 张,其中包含行人、机动车、自行车等交通目标 2101 个,对数据集采用不同算法进行实验比较,模型检测精度和速度对比如下表 1 所示,实验证明,经过图像融合和改进后的模型分类精度更高,检测速度更快。

表 1 不同算法对比结果

Tab. 1 Comparison results of different methods

目标检测算法	图像	精度	速度/($f \cdot s^{-1}$)
Faster RCNN	可见光图像	0. 7023	80
Faster RCNN	红外图像	0. 8076	90
Faster RCNN	可见光图像与红外图像融合	0. 9078	98
本文改进 Faster RCNN	可见光图像与红外图像融合	0. 9449	140

5 结 论

本文在研究典型的深度卷积神经网络算法用于行人目标检测时,以 Faster RCNN 算法为基础,采用空间金字塔池化网络、特征融合方式改进了网络中的函数,有效提高了汽车驾驶场景中应对环境条件差、目标距离汽车远近的尺度问题带来的目标检测的准确率、提高了锚点选择框在神经网络中的算法效率。理论分析和计算机程序实验数据可以说明,在道路中借助于红外图像,改进后的深度神经网络在行人检测中提高了有效性。因此,在汽车驾驶场景应用中,利用本算法可以更有效地检测行人目标。

参考文献:

[1] Mueller A S, Cicchino J B, Zuby D S. What humanlike errors do autonomous vehicles need to avoid to maximize safety? [J]. Journal of Safety Research, 2020, 75: 310 – 318.

[2] Fagnant D J, Kockelman K. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations[J]. Transportation Research Part A: Policy and Practice, 2015, 77: 167 – 181.

[3] Dollár Piotr, Wojek Christian, Schiele Bernt, et al. Pedestrian detection: an evaluation of the state of the art[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34: 743 – 61.

[4] Li Zongmin, Xing Minmin, Liu Yujie, et al. Pedestrian target detection using faster RCNN and similarity measurement[J]. Journal of Graphics, 2018, 39(5): 901 – 908(in Chinese)
李宗民,邢敏敏,刘玉杰,等. 结合 Faster RCNN 和相似性度量的行人目标检测[J]. 图学学报,2018,39(5): 901 – 908.

[5] Wang Maoqi. Research on pedestrian target detection and tracking technology based on YOLO algorithm[D]. Nanjing: Nanjing University of Science and Technology, 2021 (in Chinese)
王茂琦. 基于 YOLO 算法的行人目标检测与跟踪技术研究[D]. 南京:南京理工大学,2021.

[6] Zhao Jiuxiao, Liu Yi, Li Guoyan. Video pedestrian object detection based on improved SSD[J]. Sensors and Microsystems, 2022, 41(1): 146 – 149, 1561(in Chinese)
赵九霄,刘毅,李国燕. 基于改进 SSD 的视频行人目标检测[J]. 传感器与微系统, 2022, 41(1): 146 – 149, 1561.

[7] Zhang Ruzhen. Infrared target detection and recognition based on deep learning[D]. Beijing: University of Chinese Academy of Sciences. 2021(in Chinese)
张汝榛. 基于深度学习的红外目标检测识别[D]. 北京:中国科学院大学,2021.